日　本　国　特　許　庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２００３年　１月　９日

出願番号
Application Number: 特願２００３－００２９８１

[ST.10/C]: [ＪＰ２００３－００２９８１]

出願人
Applicant(s): 沖電気工業株式会社

２００３年　４月　１日

特許庁長官
Commissioner,
Japan Patent Office 太田信一郎

出証番号　出証特２００３－３０２２４４７

317 59-190543

【書類名】　特許願
【整理番号】　KN002566
【提出日】　平成15年 1月 9日
【あて先】　特許庁長官　太田　信一郎　殿
【国際特許分類】　G06F 15/40
【発明者】
　【住所又は居所】　東京都港区虎ノ門１丁目７番１２号　沖電気工業株式会社内
　【氏名】　池野　篤司
【特許出願人】
　【識別番号】　000000295
　【氏名又は名称】　沖電気工業株式会社
　【代表者】　篠塚　勝正
【代理人】
　【識別番号】　100090620
　【弁理士】
　【氏名又は名称】　工藤　宣幸
【先の出願に基づく優先権主張】
　【出願番号】　特願2002-187698
　【出願日】　平成14年 6月27日
【手数料の表示】
　【予納台帳番号】　013664
　【納付金額】　21,000円
【提出物件の目録】
　【物件名】　明細書　1
　【物件名】　図面　1
　【物件名】　要約書　1
　【包括委任状番号】　9006358
【プルーフの要否】　要

【書類名】　　明細書

【発明の名称】　情報区分装置、方法及びプログラム、並びに、情報区分プログラムを記録した記録媒体

【特許請求の範囲】

【請求項１】　入力された電子文書中の情報を区分する情報区分装置において、

分割行に表れ得る所定の文字列を規定する１又は複数の分割パターンを記憶している分割パターン記憶手段と、

入力された上記電子文書を上記分割パターン記憶手段に記憶されている上記分割パターンと照合して、上記電子文書を、複数の部分文書に分割する文書分割手段とを備える

ことを特徴とする情報区分装置。

【請求項２】　分類を特定し得る所定の文字列を規定する、分類情報が付与されている複数のラベリングパターンを記憶しているラベリングパターン記憶手段と、

上記文書分割手段によって分割された上記各部分文書をそれぞれ、上記ラベリングパターン記憶手段に記憶されている上記ラベリングパターンと照合して、分類情報を付与するラベリング手段とをさらに備える

ことを特徴とする請求項１に記載の情報区分装置。

【請求項３】　入力された上記電子文書の種類を判別する文書種類判別手段をさらに備え、

上記文書分割手段が、判別された文書種類用の上記分割パターンを用いて部分文書への分割を行うと共に、

上記ラベリング手段が、判別された文書種類用の上記ラベリングパターンを用いて分類情報の付与を行う

ことを特徴とする請求項２に記載の情報区分装置。

【請求項４】　入力された上記電子文書における、同様な文字列を同様な位置に含む複数行の存在を認識して、上記分割パターンを生成し、上記分割パター

ン記憶手段に登録する分割パターン生成手段をさらに備えることを特徴とする請求項１～３のいずれかに記載の情報区分装置。

【請求項５】　入力された電子文書中の情報を区分する情報区分方法において、

入力された上記電子文書を、分割行に表れ得る所定の文字列を規定する分割パターンと照合して、上記電子文書を、複数の部分文書に分割する文書分割工程を有する

ことを特徴とする情報区分方法。

【請求項６】　上記文書分割工程によって分割された上記各部分文書をそれぞれ、分類を特定し得る所定の文字列を規定する、分類情報が付与されているラベリングパターンと照合して、分類情報を付与するラベリング工程をさらに有することを特徴とする請求項５に記載の情報区分方法。

【請求項７】　入力された上記電子文書の種類を判別する文書種類判別工程をさらに有し、

上記文書分割工程が、判別された文書種類用の上記分割パターンを用いて部分文書への分割を行うと共に、

上記ラベリング工程が、判別された文書種類用の上記ラベリングパターンを用いて分類情報の付与を行う

ことを特徴とする請求項６に記載の情報区分方法。

【請求項８】　入力された上記電子文書における、同様な文字列を同様な位置に含む複数行の存在を認識して、上記分割パターンを生成して登録する分割パターン生成工程をさらに有することを特徴とする請求項５～７のいずれかに記載の情報区分方法。

【請求項９】　請求項５～７のいずれかに記載の情報区分方法の各工程をコンピュータが処理し得るコードで記述したことを特徴とする情報区分プログラム。

【請求項１０】　請求項９の情報区分プログラムを記録していることを特徴とする記録媒体。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、情報区分装置、方法及びプログラム、並びに、情報区分プログラムを記録した記録媒体に関し、特に、複数の情報が記載されている電子文書内の情報を分割して分類しようとするものである。

【０００２】

【従来の技術】

近年、インターネット等のネットワーク技術の普及により国内外の大量の電子文書へのアクセスが可能になり、大量の電子文書情報を分類する等の知的作業の自動化の必要性が高まってきている。

【０００３】

昨今発展を見せている電子文書の入手方法の一つに、メールマガジン（メールによる雑誌・新聞に類したもの）があげられる。これは、購読希望者に複数の情報をまとめて一つの電子メールに記載して送るというものである。

【０００４】

このような電子メールは、複数の情報を記載した電子文書と見なすことができ、その情報を分類するためには電子文書内の各情報を適切に分割してやる必要がある。

【０００５】

【特許文献１】　　特開２０００－２８５１４０号公報

特許文献１には、文書データの構造情報（ＨＴＭＬのタグや文字のフォント情報など）を基に文書データを分割する手段や、文書要素（例えば単語）や要素付随情報（例えば品詞）を基に文書データを分割する手段を設けることにより、情報の分類の一助としている装置の例が示されている。

【０００６】

【発明が解決しようとする課題】

しかしながら、上記公報記載の装置では、メールマガジンのように明確な構造情報を持っていない電子文書には適用できないという問題がある。

【０００７】

また、仮に、あるメールマガジンを適切に分割する情報を指定したとしても、複数のメールマガジンを受け取っている場合、各々が異なる種類の分割情報（分割パターン）を必要とする可能性が高く、メールマガジンの種類によっては、適切な分割パターンを選択して分割することができないという課題がある。

【０００８】

さらに、受け取るメールマガジンが増加すれば、分割パターンの種類も増加するが、それらを人手で指定するのは手間がかかるという課題がある。

【０００９】

そのため、メールマガジン等のような明確な構造情報を持っていない電子文書の各情報を、適切に分割できる情報区分装置等が望まれている。

【００１０】

【課題を解決するための手段】

第１の本発明の情報区分装置は、入力された電子文書中の情報を区分するものであって、（１）分割行に表れ得る所定の文字列を規定する１又は複数の分割パターンを記憶している分割パターン記憶手段と、（２）入力された上記電子文書を上記分割パターン記憶手段に記憶されている上記分割パターンと照合して、上記電子文書を、複数の部分文書に分割する文書分割手段とを備えることを特徴とする。

【００１１】

第２の本発明の情報区分方法は、入力された電子文書中の情報を区分するものであって、入力された上記電子文書を、分割行に表れ得る所定の文字列を規定する分割パターンと照合して、上記電子文書を、複数の部分文書に分割する文書分割工程を有することを特徴とする。

【００１２】

第３の本発明の情報区分プログラムは、第２の本発明の情報区分方法の工程をコンピュータが処理し得るコードで記述したことを特徴とする。

【００１３】

第４の本発明の記録媒体は、第３の本発明の情報区分プログラムを記録していることを特徴とする。

【００１４】

【発明の実施の形態】

（Ａ）第１の実施形態

以下、本発明による情報区分装置、方法及びプログラム、並びに、情報区分プログラムを記録した記録媒体の第１の実施形態を図面を参照しながら詳述する。

【００１５】

（Ａ－１）第１の実施形態の構成

図１は、第１の実施形態の情報区分装置の機能的構成を示すブロック図である。例えば、第１の実施形態の情報区分装置は、通信機能を有するパソコン等の情報処理装置に対し、ＣＤ－ＲＯＭやフロッピー（登録商標）ディスク等の記録媒体に記録されている情報区分プログラムをインストールすることで実現されるが、機能的には、図１で表すことができる。

【００１６】

図１において、第１の実施形態の情報区分装置は、文書種類判別部１と、文書分割部２と、ラベリング部３と、判別パターンデータ記憶部４と、分割パターンデータ記憶部５と、ラベリングパターンデータ記憶部６とを有する。

【００１７】

文書種類判別部１は、判別パターンデータ記憶部４の判別パターンデータを参照して、適用すべき分割パターンとラベリングパターンを決定するために、入力された電子文書（適宜、文書と呼ぶ）の種類を判別するものである。

【００１８】

なお、この第１の実施形態では、複数の全く異なる情報が一つの電子文書内に含まれている電子文書（例えばニュースのメールマガジン）を入力対象としているものである。また、構造情報は持たないが、人間が簡単に認識できるように記号などの表層情報を用いて明示的に内容の区切りが記述されている電子文書を入力対象としているものである。

【００１９】

文書分割部２は、文書種類判別部１の判別結果（すなわち、電子文書の種類）により決定された、分割パターンデータ記憶部５中の分割パターンデータを適用

して、入力された電子文書を分割するものである。

【００２０】

ラベリング部３は、文書種類判別部１の判別結果（すなわち、電子文書の種類）結果により決定された、ラベリングパターンデータ記憶部６中のラベリングパターンデータを適用して、文書分割部２により分割された入力文書の各部分に対してラベリングを行なうものである。

【００２１】

判別パターンデータ記憶部４に記憶されている判別パターンデータは、文書種類判別部１が電子文書の種別を判別するためのデータの集合である。最も単純な形式の判別パターンとしては、特定の文字列（例えば、メールマガジンであれば、メールマガジンのタイトルやＩＤ番号）が挙げられる。

【００２２】

図２は、判別パターンデータの一例を示している。各レコードは、文書種類と、その文書種類に適用する判別パターンとを含んでいる。図２に示すように、ある種類の電子文書に対し、複数の判別パターンデータが存在していても構わない。

【００２３】

分割パターンデータ記憶部５に記憶されている分割パターンデータは、文書分割部２が電子文書を分割するためのデータであり、例えば、図３に示すような文書種類と分割パターンとを対応付けたデータである。図３の分割パターンは、正規表現で記載されているので、パターン中の記号「＾」は「行頭」、「．」は「任意の一文字」、「＊」は「直前の文字が０回以上出現する」ことを意味している。例えば、図３における「＾＝＝＝＝．＊」は、「行頭から半角のイコール記号『＝』が４回出現した後にある文字が０回以上出現する」というパターンを示していることになる。図３に示すように、ある種類の電子文書に対し、複数の分割パターンデータが存在していても構わない。また、電子文書の種類を問わずに適用する分割パターンデータを設けていても良い。

【００２４】

ラベリングパターンデータ記憶部６に記憶されているラベリングパターンデー

タは、文書分割部２が分割した電子文書の各部分（各情報）に対して、ラベリング部３が分類情報を付与する（ラベリングを行なう）ためのデータであり、図４に示すような、文書種類と、ラベリングパターンと、ラベル名とを対応付けたデータの集合である。図４に示すラベリングパターンも、正規表現で記載されている。図４に示すように、ある種類の電子文書に対し、通常、複数のラベリングパターンデータが存在する。また、電子文書の種類を問わずに適用するラベリングパターンデータを設けていても良い。

【００２５】

（Ａ－２）第１の実施形態の動作

以下、第１の実施形態の情報区分装置の動作（情報区分方法）を、各構成要素１～３毎の動作で説明する。

【００２６】

まず、文書種類判別部１の動作を説明する。

【００２７】

文書種類判別部１は、判別パターンデータ記憶部４に記憶されている各パターンデータを用いて、入力された電子文書内をパターンマッチさせることにより文書種類を判別する。なお、入力文書は、ネットワークを介して取り込んでも良く、記憶媒体から取り出しても良く、その入力方法は任意である。

【００２８】

ここで、入力文書が図５に示すような電子文書であった場合には、図２における第１番目や第２番目のパターンデータの存在により、図５の電子文書は「ビジネスメールマガジン１」という種別であると判別される。

【００２９】

なお、複数のパターンデータがマッチし、かつ、その判別結果が矛盾する場合には、多数決（マッチ数が多いもの）により決定したり、矛盾が生じる旨をユーザに通知するなどの機能を設けても良い。

【００３０】

次に、文書分割部２の動作を説明する。

【００３１】

文書分割部２は、上述したように、分割パターンデータ記憶部５に記憶されている、判別された文書種類の各分割パターンデータを用いて、入力された電子文書を複数の部分文書（情報）に分割する。

【００３２】

図５の電子文書が、文書種類判別部１によって「ビジネスメールマガジン１」という種別と判別されたので、図３の第１番目及び第２番目の分割パターンが適用可能である。すなわち、（１）先頭から「－」（半角のハイフン）が一定数以上連続している、（２）先頭から「＝」（半角の等号）が一定数以上連続している、の部分が分割パターンとなるので、その位置（行）で入力文書を部分文書（情報）に分割する。

【００３３】

分割後の各部分文書は、データ全般を記憶している記憶装置上に元データとは別に記憶されることになる。なお、各部分文書の記憶部は、文書分割部２に含まれているように、図１では示している。

【００３４】

また、分割に用いた分割パターンそのものは、（１）分割後の部分文書には含めない（分割パターンは削除される）、（２）分割位置の前後の部分文書のいずれかに含める、（３）分割位置の前後の両方の部分文書に含める（パターンは複製される）、のいずれかの方法を適用する。

【００３５】

分割パターンの取扱いについて（２）の方法を適用した場合には、図５の入力文書は、図６に示すような５個の部分文書に分割される。

【００３６】

次に、ラベリング部３の動作を説明する。

【００３７】

ラベリング部３は、上述したように、ラベリングパターンデータ記憶部６に記憶されている、判別された文書種類の各ラベリングパターンデータを用いて、パターンがマッチした部分文書をラベリングする。

【００３８】

図５（図６）の電子文書が文書種類判別部１によって「ビジネスメールマガジン１」という種別と判別されたので、図４の第１番目～第４番目のラベリングパターンデータが利用され、その結果、部分文書１に対して「広告」、部分文書２に対して「タイトル」、部分文書３及び４に対して「記事本文」、部分文書５に対して「注釈」のようにラベリングされる。

【００３９】

例えば、部分文書１には、「－－－ＰＲ－」というパターンが存在するので、図４の第２番目の行が適用され、「広告」とラベリングされる。これらのラベル情報は、各部分文書と組にして保持される。

【００４０】

ラベル情報を有する部分文書の情報は、ユーザの操作等に応じて、表示出力されたり、印刷出力されたり、他へ送信されたりする。この際、ユーザは、例えば、記事本文だけを指定して出力させたりすることもできる。また、ラベル情報を有する部分文書の情報は、さらなる加工処理が実行されても良い。例えば、記事本文に対して要約作成処理を施すようにしても良い。

【００４１】

（Ａ－３）第１の実施形態の効果

以上のように、第１の実施形態によれば、簡単なパターンによる分割パターンデータやラベリングパターンデータを用意するだけで、ＸＭＬやＨＴＭＬやＳＧＭＬ等で記述されたような明確な構造を持つ電子文書ではなくても、電子文書を分割して分類することができる。

【００４２】

しかも、文書種類判別部を設けたので、複数の分割パターンを管理しておき、様々な種類の電子文書を対象に電子文書を分割して分類することができる。

【００４３】

（Ｂ）第２の実施形態

次に、本発明による情報区分装置、方法及びプログラム、並びに、情報区分プログラムを記録した記録媒体の第２の実施形態を図面を参照しながら詳述する。

【００４４】

（Ｂ－１）第２の実施形態の構成

図７は、第２の実施形態の情報区分装置の機能的構成を示すブロック図であり、第１の実施形態に係る図１との同一、対応部分には、同一符号を付して示している。

【００４５】

第２の実施形態の情報区分装置は、第１の実施形態の構成に、分割パターン生成部７を付加した構成となっている。

【００４６】

分割パターン生成部７は、入力された電子文書を基に分割パターンを生成するものである。分割パターン生成部７によって生成された分割パターンは、文書分類判別部１によって判別された文書種類に対応付けられ、分割パターンデータとして分割パターンデータ記憶部５に記憶される。

【００４７】

分割パターン生成部７以外の部分は、第１の実施形態と同様の機能を担っているので、その説明は省略する。

【００４８】

（Ｂ－２）第２の実施形態の動作

第１の実施形態と動作が異なるのは分割パターン生成部７の動作だけなので、以下では、その動作のみを、図８のフローチャートを参照しながら説明する。

【００４９】

分割パターン生成部７は、入力文書が与えられると、入力文書を行ごとに分割する（ステップ８０１）。次に、先頭から所定番目（例えば３０番目）の文字の全てが一致する行のグループを作ると共に、その行グループに属する行数も計数しておく（ステップ８０２）。

【００５０】

例えば、上述した図５の電子文書が入力文書である場合、ステップ８０２の処理を終えた段階では、図９に示すような行グループが作成される。

【００５１】

その後、分割パターン生成部７は、複数のメンバ（行）（ここでは２以上とす

る）を持つ行グループのみを選択してパターン記述を行う（ステップ８０３）。最も簡単なパターン記述法は文字列そのものであるが、必要に応じて正規表現などに書き改めるなどの手法を用いても良く、文書分割部２が理解できる形式を出力するものであれば特に手法は問わない。

【００５２】

その後、分割パターン生成部７は、文書種類判別部１から、文書種類のデータを取り込んで分割パターンデータを完成させて分割パターンデータ記憶部５に登録する（ステップ８０４）。なお、文書種類のデータを含まない分割パターンデータを登録するようにしても良い。

【００５３】

上述したステップ８０２の行一致を判別するための文字数や、ステップ８０３の登録に値するかを判別するためのメンバ（行）数は自由に設定しても良い。また、ステップ８０２において「先頭から複数文字」としているが、「末尾から」であっても良く、「先頭および末尾から」であっても良く、「先頭や末尾に関係なく」であっても良い。また、それらを自由に設定できる形式であっても良い。

【００５４】

（Ｂ－３）第２の実施形態の効果

第２の実施形態によれば、第１の実施形態と同様な効果を奏すると共に、さらに、自動的に分割パターンデータを生成して登録することができるという効果をも奏する。

【００５５】

（Ｃ）他の実施形態

上記各実施形態においては、入力文書の分割を行った後に、各部分文書に対するラベリングを行うものを示したが、入力文書の分割及び分割された各部分文書に対するラベリングを並行して同時に行なっても良い。

【００５６】

また、分割パターンデータをラベリングパターンデータの一部として用いるようにしても良い。

【００５７】

上記各実施形態は、入力文書が横書き文書であるものを示したが、縦書き文書に対応できるようにしても良い。この場合、縦方向の行パターンを利用して、上記各実施形態と同様な処理を行うようにすれば良い。

【００５８】

また、上記各実施形態では、文書種類判別部が入力文書の種類を自動判別するものを示したが、ユーザ等が入力文書の種類を入力するものであっても良い。また、全ての分割パターンやラベリングパターンを、文書種類に関係なく、登録しておき、入力文書の種類を特定することなく、部分文書への分割、及び、分割された部分文書へのラベリングを行うようにしても良い。さらに、ある種類の入力文書専用の情報区分装置として装置を構成しても良い。

【００５９】

さらに、上記各実施形態の分割パターンは、その行が分割行であることを確定するものであったが、ある分割パターン（様子見分割パターン）に一致する行より所定行以内に、他の分割パターンに一致する行がないことを判明した場合に、分割行と確定するような分割パターン（様子見分割パターン）を設けるようにしても良い。

【００６０】

【発明の効果】

以上のように、本発明によれば、メールマガジン等のような明確な構造情報を持っていない電子文書の各情報を、適切に分割することができる。

【図面の簡単な説明】

【図１】

第１の実施形態の情報区分装置の機能的構成を示すブロック図である。

【図２】

第１の実施形態の判別パターンデータ例を示す説明図である。

【図３】

第１の実施形態の分割パターンデータ例を示す説明図である。

【図４】

第１の実施形態のラベリングパターンデータ例を示す説明図である。

【図５】

第１の実施形態の動作説明に適用する入力文書例を示す説明図である。

【図６】

図５の入力文書に対する文書分割処理後のデータを示す説明図である。

【図７】

第２の実施形態の情報区分装置の機能的構成を示すブロック図である。

【図８】

第２の実施形態の分割パターン生成部の動作を示すフローチャートである。

【図９】

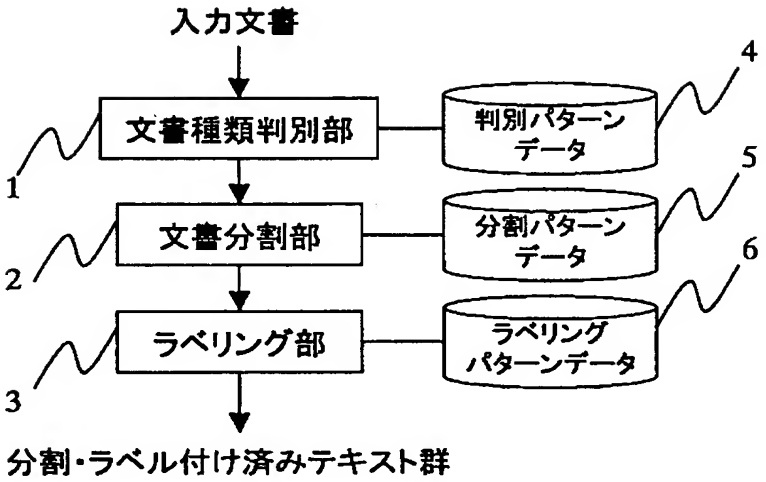
第２の実施形態の分割パターン生成時における入力文字のグループ化の説明図である。

【符号の説明】

１…文書種類判別部、２…文書分割部、３…ラベリング部、４…判別パターンデータ記憶部、５…分割パターンデータ記憶部、６…ラベリングパターンデータ記憶部、７…分割パターン生成部。

【書類名】　　図面

【図１】



【図２】

文書種類	判別パターン
ビジネスメールマガジン1	“DEFビジネスメールマガジン”
ビジネスメールマガジン1	ID=0000111
ビジネスメールマガジン2	“XYZニュース”
ビジネスメールマガジン2	MailMagazineID=999

【図３】

文書種類	分割パターン
ビジネスメールマガジン1	^====.*
ビジネスメールマガジン1	^----.*
ビジネスメールマガジン2	.....*
ビジネスメールマガジン2	^=-=-.*

【図４】

文書種類	ラベリングパターン	ラベル
ビジネスメールマガジン1	^●.*	記事本文
ビジネスメールマガジン1	---PR-	広告
ビジネスメールマガジン1	ID=000.*	タイトル
ビジネスメールマガジン1	^Copyright.*	注釈
ビジネスメールマガジン2	^◆.*	記事本文
ビジネスメールマガジン2	^©.*	注釈

【図５】

---PR----------
ABC Computer 新製品発売中！！

DEF ビジネスメールマガジン 朝刊
ID-0000111

●GHI株式会社、JKFコーポレーションとの提携を発表
http://www.aaa.bbb/ccc/ddd1234.html

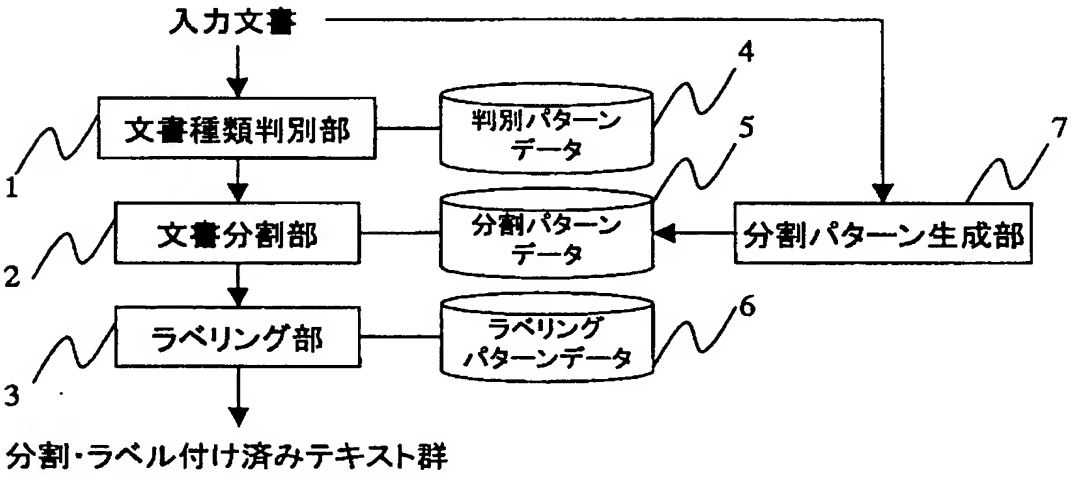
GHI株式会社は××日、JKFコーポレーションと＊＊分野での提携に合意したと発表した。・・・
==========
●株式会社LMN、第1四半期の売上高は○○億円
http://www.aaa.bbb/ccc/ddd5678.html

LMNは△△の好調な売れ行きを反映して、第1四半期の売上高は○○億円と前期比＊＊%増を・・・
==========
Copyright 2002 OPQ社　無断転載を禁ず

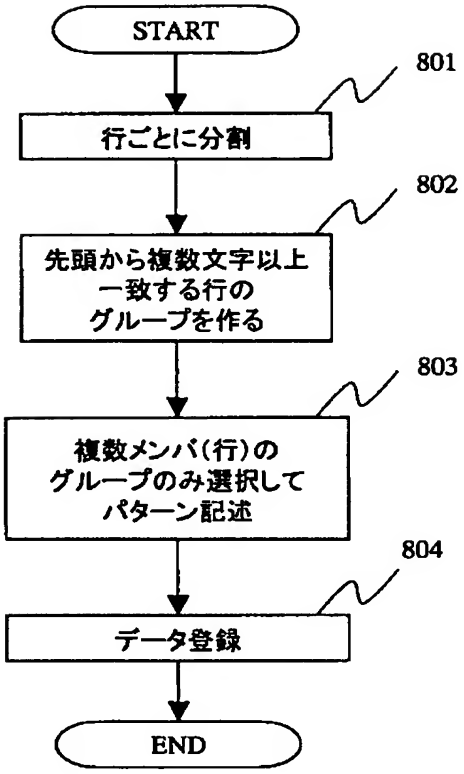
【図６】

部分文書1	---PR---------- ABC Computer 新製品発売中！！
部分文書2	---------- DEF ビジネスメールマガジン 朝刊 ID-0000111
部分文書3	---------- ●GHI株式会社、JKFコーポレーションとの提携を発表 http://www.aaa.bbb/ccc/ddd1234.html GHI株式会社は××日、JKFコーポレーションと＊＊分野での提携に合意したと発表した。・・・
部分文書4	========== ●株式会社LMN、第1四半期の売上高は〇〇億円 http://www.aaa.bbb/ccc/ddd5678.html LMNは△△の好調な売れ行きを反映して、第1四半期の売上高は〇〇億円と前期比＊＊%増を・・・
部分文書5	========== Copyright 2002 OPQ社　無断転載を禁ず

【図７】



【図８】



【図９】

先頭からの文字	行数
----(以下30個目まで)	2
==== (以下30個目まで)	2
---PR--- (以下30個目まで)	1
・・・	・・・
http://www.aaa.bbb/ccc/ddd1234	1
・・・	・・・

【書類名】　　要約書

【要約】

【課題】　メールマガジン等のような明確な構造情報を持っていない電子文書の各情報を、適切に分割する。

【解決手段】　本発明では、入力された電子文書を、分割行に表れ得る所定の文字列を規定する分割パターンと照合して、複数の部分文書に分割する。その後、分割された各部分文書をそれぞれ、分類を特定し得る所定の文字列を規定する、分類情報が付与されているラベリングパターンと照合して、分類情報を付与することが好ましい。

【選択図】　　図１

出　願　人　履　歴　情　報

識別番号　　〔０００００２９５〕

１．変更年月日　　１９９０年　８月２２日
［変更理由］　新規登録
住　所　　東京都港区虎ノ門１丁目７番１２号
氏　名　　沖電気工業株式会社